CSC 4740/6740 Data Mining

Assignment 1

Due Date: 11:59 pm, Thursday, September 22, 2022

Note: Even though these statistics are simple to compute, which can be done manually, I suggest that you calculate them using programs, either Python or Matlab are recommended. You can call the API functions from any libraries. This will be better for you to get familiar with these API functions. Real-life dataset will be large and computer programs are needed.

1. (10 points) Suppose we have the BestBuy customer data in the following table.

Customer	**Age**
David	46
Lisa	25
Michael	27
Susan	27
William	28
Mat	36
James	53
Kevin	27
Paul	18
Anthony	25

1.1) Please calculate the mean, median, and mode.

2. (25 points) Suppose we have the climate data for Atlanta in the following table.

Climate data for Atlanta

Month	**Temperature (℉)**
Jan	52.3
Feb	56.6
Mar	64.6
Apr	72.5
May	79.9
Jun	86.4
Jul	89.1
Aug	88.1
Sep	82.2
Oct	72.7
Nov	63.6
Dec	54.0

2.1) Please compute the five-number summary of this dataset.

2.2) Will there be outliers if we use boxplot to visualize the five-number summary? If yes, please indicate which data objects are outliers. Please briefly explain your answers.

2.3) Please visualize the data by using plot function in Matlab or some similar functions in other software. You can use any software. Based on the plotted curve, please also briefly describe the visualization result.

3. (15 points) Suppose we have the customers' information in the following table.

Customer	**David**	**Susan**	**Lisa**
Profession	Manager	Manager	Programmer
Education	B.Sc.	B.Sc.	M.Sc.
Hobbies	Golf	Swimming	Swimming

3.1) Which types of attributes are there in the table?

3.2) Please compute the similarity values between "David" and "Susan".

3.3) Please compute the similarity values between "Susan" and "Lisa".

4. (15 points) Suppose we have the patients' information in the following table.

Patient	**Tom**	**Mat**	**Lucy**
Fever	Yes	No	Yes
Cough	No	Yes	Yes
Sleepy	Yes	No	No
Headache	Yes	Yes	No
Running nose	Yes	Yes	No
Fatigue	Yes	Yes	Yes
Sweaty	Yes	No	Yes
Dizziness	Yes	Yes	Yes

4.1) Which types of attributes are there in the table?

4.2) Compute the similarity values between "Tom" and "Mat";

4.3) Compute the similarity values between "Mat" and "Lucy".

5. (15 points) Suppose we have the Fisher's iris data in the following table.

Flower	A	B	C
Sepal Length	5.1	7.0	4.8
Sepal Width	3.5	3.2	3.4
Petal Length	1.4	4.7	1.9
Petal Width	0.2	1.4	0.2

Please choose one similarity measure and solve the following problems.

5.1) Which types of attributes are there in the table?

5.2) Which type of similarity measure do you choose?

5.3) Compute the similarity values between "A" and "B";

5.4) Compute the similarity values between "B" and "C".

6. (15 points) Suppose we have the customer information in the loan company in the following table.

Customer	Kevin	John	Daniel
Credit Score Range	Excellent	Very good	Good
Salary Range	High	Very High	Medium
Age	Senior	Middle Age	Young

The ranking options within each attribute are provided in the following tables.

Credit Score Range
Excellent
Very good
Good
Fair
Poor

Salary Range
Very High
High
Medium
Low

Age
Senior
Middle Age
Young

6.1) Which types of attributes are there in the table?

6.1) Compute the similarity values between "Kevin" and "John".

6.2) Compute the similarity values between "John" and "Daniel".

7. (5 points) Please normalize the following dataset by using the min-max normalization method. The new range should be [0, 1].

Patient	Tom	Mat	Lucy	Brian
Height (feet)	5.7	6.2	5.1	6.4